Flight Delay Prediction Based on Gradient Boosting Ensemble Models

Rahemeen Khan, Salas Akbar and Tooba Zahid

November 18, 2022

Flight Delay Prediction Based on Gradient Boosting Ensemble Techniques

Rahemeen Khan
Department of Software Engineering
Bahria University
Karachi, Pakistan
rahemeen.bukc@bahria.edu.pk

Salas Akbar
Department of Computer Science
Bahria University
Karachi, Pakistan
salasakbar.bukc@bahria.edu.pk

Tooba Ali Zahed
Department of Computer Science
Bahria University
Karachi, Pakistan
toobazahid.bukc@bahria.edu.pk

Abstract— **In recent years, the volume of airline transportation has increased with the rapid development of aviation. With an increased demand for flights, aviation is confronted with the issue of flight delays, which becomes a series of issues that must be addressed efficiently. Correct flight delay prediction can improve airport operations efficiency and passenger travel comfort. The current study uses Gradient boosting ensemble models to build a machine learning flight delay prediction model. The Airline dataset was subjected to three different gradient boosting techniques: CatBoost, LightGBM, XGBoost, and Decision tree. To validate the performance and efficiency of the proposed method, a comparative analysis between the top performed Boosting techniques with other Ensemble Techniques is performed. CatBoost improves prediction accuracy while maintaining stability, according to the comparison results on the given dataset.**

Keywords—Airline, Flight Delay Prediction, Ensemble Learning, CatBoost, LightGBM, XGBoost, Decision tree, Boosting Techniques

I. Introduction

Flight operations on commercial aircraft have become increasingly complex and dynamic. The daily operations of airlines require adjustments to be made in response to factors such as weather conditions, mechanical issues, or passenger complaints. These variables can also impact routes and schedules, increasing variability in flight activity at commercial airports. Managing the complex interaction between passengers, planes, airports and the demands of aviation stakeholders is challenging for airlines and traffic flow managers at commercial airports who must respond quickly to unexpected changes in demand.

Delays and their possible repercussions are an inevitable part of operating an airport. Airlines, passengers, and traffic managers all have a direct stake in reducing these delays to as low a level as possible. Accurate delay prediction models are needed to ensure efficient airport operations and reduce costs to passengers.

The goal of this research is to inspect the impact of various flight delays on airport on-time performance and airline operations as well as to examine how these variables are affected by unobserved heterogeneity in data. More specifically, this study used a two-step model approach which examined the positive and negative effects of significant characteristics or factors on flights that experienced delays.

As it pertains to the prediction aspect, studies which incorporate ML models in flight delay analysis tend to overlook the potential improvements of prediction performance through proper exploratory data analysis and hyperparameter tuning. On account of such, while providing a deeper examination of models' outcomes, it also examines the application of a metaheuristic algorithm in hyperparameter tuning of the ML models for delay prediction [1]. Furthermore, this research concluded that most of the variables present in the dataset have low influence on the flight delays that occur. However, there are few Airlines whose delay rates are much higher than their competitors.

II. Literature Review

Many studies have been undertaken to anticipate flight delays using various machine learning and deep learning methodologies. Decision tree tends to work more accurately in delay prediction, with an accuracy score of 93%, compared to logistic regression and neural network, which have accuracy scores of 92% and 91%, respectively [1], demonstrating that classification models outperform deep learning models over the same nature of the data.

According to another study[2], Support Vector Machine (SVM) has a Mean Average Error (MAE) of 7.84% compared to K-Nearest Neighbors (KNN) and Random Forest, which have MAEs of 27.62% and 31.14%, respectively, demonstrating that ensemble classification models perform better for delay prediction.

In latest research [3], Random Forest performance examined on a cluster environment using an airline dataset, this model predicted flight delays with an accuracy of 92.7%.

In 2021, a research study based on the stacking of KNN, Random Forest, Logistic regression, Decision Tree, and Gaussian Naive Bayes was utilized to forecast flight delays at Boston Logan International Airport, and the stacking model surpassed baseline techniques with an accuracy of 82.2% [4].

Another research study was conducted in which a proposed model developed to find out the significant variable impact on the flight delay by implementing the mixed logit model and exploring the non-linear relationship with the help of SVM trained by the Artificial Bee Colony algorithm on the Miami International Airport dataset [5]. As a consequence, ABC-SVM generated 94.4% accuracy against 84.73% accuracy for the traditional SVM model.

In a recent study [6], a multiclass SVM model was used to predict and analyze flight delays. The SVM model was used to evaluate the cause and patterns that affect flight delays.

To forecast the flight delay, a logistic regression model based on Microsoft Azure cloud was used [7]. Airport data is combined with weather data to get more precise findings, and the results show that the LR model reached 80% accuracy.

In 2020, a research study employing the ensemble approach [8] generated an SVM-based model with a MEA of 9.73% that predicted flight delays better than Cat-boost (MAE: 16.348%), Bagging Regression (MAE: 14.241%), and MLP Regression (MAE: 21.956%). As a result, SVM outperforms other algorithms in ensemble techniques.

In different ensemble techniques, Ensemble Stacking (ELS) [9] results in higher accuracy of 96.44% as compared to Ensemble bagging (EBS) with accuracy of 96.14% , XGBoost and CatBoost resulted in 95.25% and 95.54% respectively.

In 2022 research [10], flight delay forecasting model is demonstratedd by integrating machine learning models based on gradient boosting, an ensemble approach, models XGBoost, LightGBM, and CatBoost, with XGBoost having a greater accuracy of 96.2%.

A model is developed to avoid airline delays using data mining and machine learning approaches, covering the top five busiest airports in the United States. The gradient boosting classifier predicted flight delays with an accuracy of 85.73% [11].

According to research, ensemble techniques are considerably more successful in forecasting flight delays than deep learning and basic machine learning-based categorization approaches. However, this assessment of the literature reveals that, except from a research study completed in 2022 [9,] there are few publications that offer comparative analysis within the various categories of ensemble approaches (boosting, bagging, Stacking). This research study compares multiple ensemble strategies to see which strategy performs best on flight delay prediction within the ensemble methodologies.

III. PROPOSED METHODOLOGY

A. *Dataset Collection*

The Airline dataset used in this research is extracted from Kaggle. The dataset contains the flight information regarding Airline, flight, Source airport, Destination airport, Day of week, Time and Delay as binary Label (0 indicating no delay in flight while 1 represents the delay).

This Airlines dataset has 539383 records and 8 different columns. TABLE I. represents the overall attributes in which 3 attributes are categorical while remaining 4 features have continuous values.

TABLE I. ATTRIBUTE STUDY

Feature Name	***Description***	***Data Type***
Airline	Types of commercial airlines	Categorical
Flight	Types of Aircraft	Continuous
Airport From	Source Airport	Categorical
Airport To	Destination Airport	Categorical
DayOfWeek	Tells you about the day of week	Continuous
Time	Time taken.	Continuous
Length	Length	Continuous

B. *Data Preprocessing*

Before implementing machine learning models, the flight dataset must be preprocessed. The following are some preprocessing techniques.

1. *Feature Encoding*

In this research, Airline dataset features such as Airline, Airport From, and Airport To are categorical values that cannot be processed directly. It must be converted into numeric values before applying traditional machine learning models. To encode categorical features, the label encoder technique is used.

2. *Data Normalization*

A feature with a similar scale can have a significant impact on the machine learning model's performance. The standardization technique is used to scale the feature values of a given dataset.

3. *Dataset Splitting*

The dataset must be split into training and testing ratios before fitting the model. The entire dataset was shuffled and split into 70% training and 30% test sets in this study.

4. *Gradient Boosting Ensemble Methods*

Based on the performance and popularity among various machine learning algorithms, gradient boosting trees (GBTs) based algorithms are well known for the structured data. The three GBT-based algorithms chosen for this study to predict flight delays are as follows.

✓ CatBoost

CatBoost, which stands for categorical boosting, was created in 2017 by the Russian company Yandex as an open-source algorithm-based tool. The CatBoost classifier is another machine learning algorithm that is effective at predicting categorical features. It is a gradient boosting implementation that uses binary decision trees as base predictors [12]. It has demonstrated superior results on categorical features when compared to other boosting models. CatBoost, in contrast to deep learning models, produces useful results even with limited training data and computational power [13].

✓ LightGBM

LightGBM is a tree-based gradient boosting model with high accuracy and fast training speed [14]. The results of multiple decision trees combine to interpret.

- ✓ XGBoost

XGBoost is an abbreviation for extreme gradient boosting. It began as a research project by Tianqi Chen in 2014 [15] and became well-known in 2016. It is a collection of decision trees built from short and simple trees. XGBoost employs the concept of parallelized implementation to improve model performance.

IV. Implementation Details

The proposed research is carried out on a Google Colaboratory notebook using Python 3.6.8 programming environment. Pandas, Numpy, sci-kit-learn, and matplotlib are the most commonly used libraries in this study.

V. Results

A. Exploratory Data Analysis

Before developing the machine-learning model, the dataset is analysed and understood using exploratory data analysis. In this study, binary classification is performed on datasets that contain 539383 observation data points with 7 features.

Fig. 1 represents the class distribution, with approximately 5500 delayed flights in this data, while approximately 10000 flights experienced no delays.

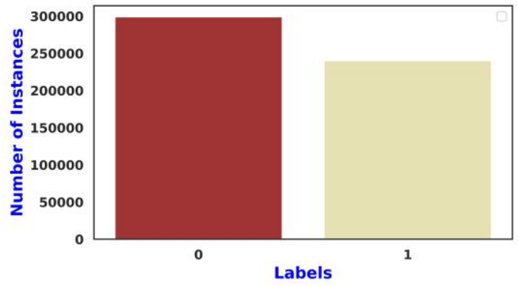

Fig. 1 Bar Visualization showing the Class Label Distribution

In this dataset, 18 different airlines flight observations are taken from 300 airports across the USA. In Fig.2 it can be seen that WestJet Airlines (WN) has the most significant number of flights while Hawaiian Airlines (HA) has the least number of flights.

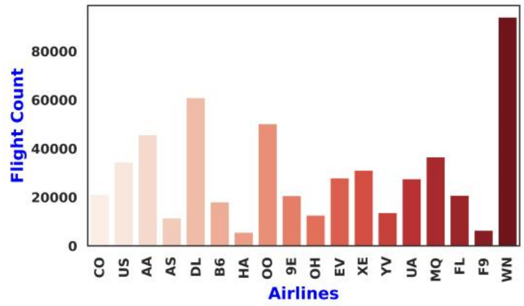

Fig. 2 Significant Airlines

This dataset contained all the major airlines and the airports along with other data such as source and destination, time taken in aviation, and flight length, to name a few. Fig. 3 analyze the ranking of airlines through the number of delays, WestJet Airlines(WN) is the most significant airline in terms of flight delay, followed by Continental Airlines(CO).

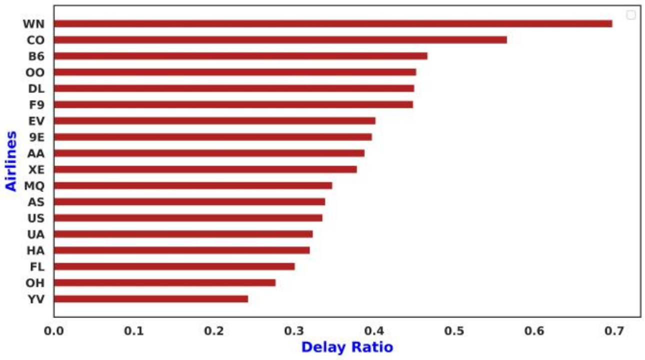

Fig. 3 Flight Delay Ratio

Fig. 4 depicts the flight density graph, which visualizes multiple airlines' arrival delay and represents the maximum distribution of flights taken on Wednesday (3) and Thursday (4) in a given week.

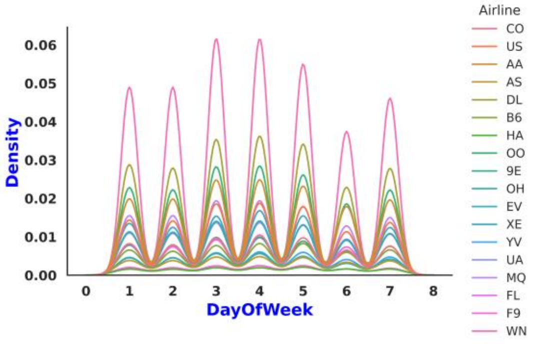

Fig. 4 Flight density graph

To find the most defining factors responsible for the delay of flights in America from the perspective of features, Fig. 5 represents the essential feature in the dataset, which was the AirportFrom variable which indicates which was the source airport, followed by the Airline itself and the flight time.

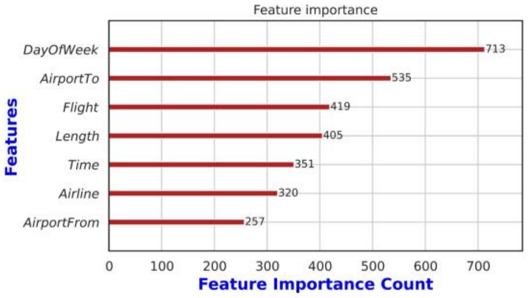

Fig. 5 Feature Importance in Flight Dataset

B. Performance Measure

The metrics used in this study to efficiently measure the performance of machine learning models are as follows.

- *Accuracy Score*

To check the model correctness for predicting the samples in the testing Set.

- *Precision*

It describes the proportion of correct predictive and overall predicted positive observations.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- *Recall*

It is defined as the fraction of correctly identified positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- *F1-Score measure*

It calculates the harmonic mean of precision and recall.

$$\text{F1} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

C. Comparative Analysis of Model Performance

Several evaluation scores are investigated to perform the comparative analysis in order to evaluate the performance of the different models. In the first stage, comparing the performance of four boosting models on the dataset reveals that CatBoost outperforms the LightGBM, XGBoost, and Decision tree models by 3%, 4%, and 4%, respectively. TABLE II shows the outcome.

TABLE II. COMPARISION BETWEEN DIFFERENT BOOSTING TECHNIQUES

Model	Accuracy	Precision	Recall	F1-score
LightGBM	0.65	0.67	0.44	0.54
XGBoost	0.64	0.67	0.40	0.50
CatBoost	**0.68**	**0.65**	**0.50**	**0.57**
Decision Tree	0.64	0.63	0.50	0.56

Finally, the proposed model's performance is compared to that of other ensemble techniques (Majority voting, Bagging and Stacking). TABLE III shows that the accuracy of the CatBoost model is higher than that of others.

Therefore, the CatBoost model has a greater tendency in forecasting the flight delay on the given complex dataset as compared to other ensemble techniques.

TABLE III. COMPARISION BETWEEN CATBOOST & OTHER ENSEMBLE TECHNIQUES

Model	Accuracy	Precision	Recall	F1-score
CatBoost	**0.68**	**0.65**	**0.50**	**0.57**
Majority voting	0.60	0.66	0.24	0.35
Bagging	0.61	0.57	0.54	0.55
Stacking	0.64	0.61	0.50	0.55

VI. RESULT DISCUSSION

The outcomes of several boosting methods applied on the said dataset are shown in Table II. In comparison to previous models, the CatBoost algorithm seems to provide more accuracy (68%). In Table III, the CatBoost algorithm is compared to various ensemble strategies; once again, CatBoost is the one with the highest accuracy, and Stacking is the second-best ensemble strategy. Determining that CatBoost , that lies in the category of Gradient Boosting, performs better when compared to other ensemble techniques as well as boosting techniques.

To improve accuracy, the CatBoost algorithm was improved a bit. The number of epochs was increased from 300 to 500, the training/testing ratio was changed from 70/30 to 60/40, and the accuracy was raised from 68% to 69%. However, additional epoch and training/testing ratio modifications resulted in a decrease in accuracy score, bringing the accuracy score down to 68%, implying that 69% accuracy is the best that can be achieved. The limited record count of the aforementioned dataset could be the cause. The accuracy of the model may improve if the dataset is used with a larger number of records. Table IV shows that accuracy has improved.

TABLE IV. COMPARISION BETWEEN CATBOOST PRERFORMANCE BASED ON PARAMETER TUNING

	Epoch	Train ratio	Test ratio	Accuracy
CatBoost *(Previous)*	300	70	30	0.68
CatBoost *(Updated)*	500	40	60	0.69

VII. CONSLUSION

Airline delays are a well-known issue in the airline business, costing important time and effort. This study examined airline data with delays to determine the most significant factors that cause aircraft delays. It investigated the performance of several ensemble strategies (boosting, bagging, majority voting, and stacking) in forecasting flight delays. Table II shows that, given the size of the dataset used in this study, the predictive CatBoost algorithm produced the most significant results with an accuracy score of 68%, followed by LighBGM with an accuracy score of 65.0%. Boosting strategies outperform other ensemble procedures, according to the findings.

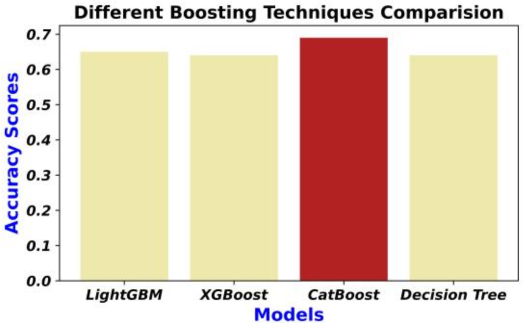

Fig. 6 Bar graph showing the Performance of different Boosting Techniques

Fig. 6 displays the accuracy score of boosting algorithms, demonstrating that CatBoost is the best performing boosting technique.

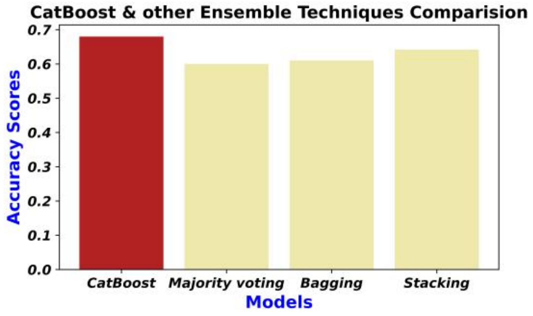

Fig. 7 Bar graph showing the Performance of CatBoost with other Ensemble Techniques

As shown in Fig. 7, CatBoost outperforms not just other boosting strategies, but also other ensemble methods. In conclusion, boosting algorithms work well on the given type of data with the different ensemble strategies, and Catboost is the best performing one among the boosting techniques.

REFERENCES

[1] Borse, Yogita & Jain, Dhruvin & Sharma, Shreyash & Vora, Aakash. (2020). Flight Delay Prediction System. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS030148.

[2] W. Wu, K. Cai, Y. Yan and Y. Li, "An Improved SVM Model for Flight Delay Prediction," 2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC), 2019, pp. 1-6, doi: 10.1109/DASC43569.2019.9081611.

[3] Cinantya, Paramita, Catur, Supriyanto, Luqman Afi, Syarifuddin, & Fauzi Adi, Rafrastara. The Use of Cluster Computing and Random Forest Algoritm for Flight Delay Prediction, International Journal of Computer Science and Information Security (IJCSIS), 2022

[4] Yi, Jia & Zhang, Honghai & Liu, Hao & Zhong, Gang & Li, Guiyi. Flight Delay Classification Prediction Based on Stacking Algorithm. Journal of Advanced Transportation. 2021.

[5] Mokhtarimousavi, Seyedmirsajad & Mehrabi, Armin. Flight delay causality: Machine learning technique in conjunction with random parameter statistical analysis. International Journal of Transportation Science and Technology(IJTST). 2022

[6] Esmaeilzadeh E, Mokhtarimousavi S. Machine Learning Approach for Flight Departure Delay Prediction and Analysis. Transportation Research Record. 2020

[7] R. Nigam and K. Govinda, "Cloud based flight delay prediction using logistic regression," International Conference on Intelligent Sustainable Systems (ICISS), 2017.

[8] X. Dou, "Flight Arrival Delay Prediction and Analysis Using Ensemble Learning," 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020, pp. 836-840, doi: 10.1109/ITNEC48623.2020.9084929.

[9] Rosalin Sahoo, Ajit Kumar Pasayat, Bhaskar Bhowmick, Kiran Fernandes & Manoj Kumar Tiwari (2022) A hybrid ensemble learning-based prediction model to minimise delay in air cargo transport using bagging and stacking, International Journal of Production Research, 60:2, 644-660, DOI: 10.1080/00207543.2021.2013563

[10] Hatıpoğlu, Irmak, et al. "Flight delay prediction based with machine learning." Logforum 18.1 (2022): 8. DOI: 10.17270/J.LOG.2022.655

[11] Chakrabarty, Navoneel. A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines. 102-107. 10.1109/IEMECONX.2019.8876970.

[12] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush & A. Gulin, CatBoost: Unbiased Boosting with Categorical Features. In Advances in Neural Information Processing Systems, 2018.

[13] Safaei N, Safaei B, Seyedekrami S, et al. E-CatBoost: An efficient machine learning framework for predicting ICU mortality using the eICU Collaborative Research Database, PLoS One. 2022.

[14] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Adv Neural Inf Process Syst, 2017.

[15] Tama BA, Im S, Lee S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. In: BioMed Research International. Hindawi 2020.